

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Физтех-школа физики и исследований им. Ландау

Машинное обучение. Конспект

Автор:

Яренков Александр Владимирович

Долгопрудный
12 июня 2024 г.

1 Введение

Определим основные понятия и покажем их на примере из таблицы под этим предложением.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

- Датасет/выборка/набор данных - данные, которые нам доступны
Вся таблица, кроме последней колонки.
- Объект/наблюдение/datum/data point - элемент выборки
Обычно(всегда ли?) мы предполагаем их независимыми и одинаково распределёнными (н.о.р.).
Строка в таблице
- Признак/feature - характеристика объекта.
Все колонки признаков образуют матрицу признаков/design matrix (X) (её столбцы x_i - вектора, имеющие все признаки объекта).
Первые 6 колонок в таблице
- Целевая (зависимая) переменная/target (y) - переменная, значение которой нас интересует
В задаче бинарной классификации (target принимает значения из множества мощностью 2) имеет специфическое название - label. Задача регрессии ($y \in \mathbb{R}$)
Жёлтый столбец в таблице
- Модель - метод обучения. Формально, действующее на признаковом пространстве отображение ($f(X)$)
- Предсказание/prediction (\hat{y}) - результат работы модели
Зелёная колонка в таблице
- Функция потерь/Loss (L) - критерий качества модели. Её значение мы пытаемся уменьшать, чтобы модель была качественнее.
Примеры функции потерь:

Mean Squared Error/дисперсия: $MSE(y, \hat{y}) = \|y - \hat{y}\|_2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \hat{y}_i)^2$

Mean Absolute Error: $MAE(y, \hat{y}) = \|y - \hat{y}\|_1 = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$

Здесь N — количество наблюдений

- Параметры - Переменные, значение которых модель выбирает сама
Например, взвешивание данных
- Гиперпараметры - параметры, которые мы фиксируем до начала работы с данными
Например, глубина решающего дерева

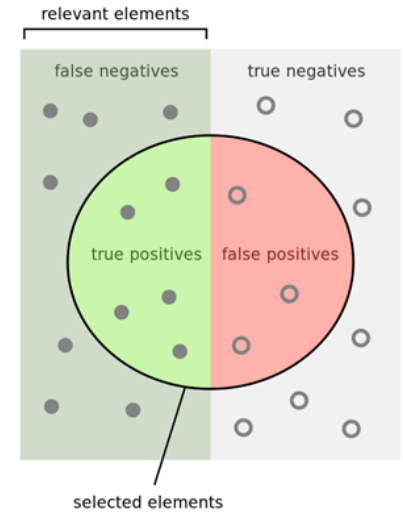
1.1 Формальная задача обучения с учителем

Обучение с учителем - обучение с наличием целевой переменной.

Поставим формально задачу обучения с учителем: $\{x_i, y_i\}_{i=1}^N$ — тренировочный датасет (training set), $f(X)$ — модель f , которой на вход подаём матрицу признаков X и получаем предсказание $f(X) = \hat{y}$, $L(X, y, f)$ — Loss function, that should be minimized

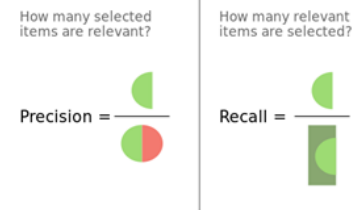
1.2 Метрики

		True condition	
Total population		Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive TP	False positive, FP Type I error
	Predicted condition negative	False negative, FN Type II error	True negative TN



Precision - доля верно классифицированных объектов (TP) ко всем классифицированным объектам (TP + FP). Она же точность

Recall - доля верно классифицированных объектов (TP) ко всем объектам этого класса (TP + FN). Т.е. мы узнаём сколько значений из класса мы НЕ потеряли.



Есть задачи в которых важнее либо precision, либо recall. Например, заболевшие люди приходят в больницу. Если им назначить лечение, не факт что им станет лучше. Но если им не назначить лечение, то им может стать сильно хуже. Поэтому нам не так важно улучшить состояние каждого (повысить precision - в данном случае доля заболевших людей, у которых улучшилось состояние), как чтобы людям не становилось ещё хуже (т.е. нужно повысить recall - в данном случае доля заболевших людей, у которых не ухудшилось состояние)

Поэтому введём понятие F_β -score. Это метрика, учитывающая важность либо precision, либо recall.

$$F_\beta = (1 + \beta^2) \frac{Precision \cdot Recall}{\beta^2 Precision + Recall}$$

Наиболее частые частные случаи:

$$F_\beta|_{\beta=0} = F_0 = 2 \cdot Precision$$

$$F_\beta|_{\beta=+\infty} = F_\infty = Recall$$

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Наивный байесовский классификатор

$\{x_i, y_i\}_{i=1}^N$ - training set. $x_i \in \mathbb{R}^p$ - вектора, $y_i \in \{C_i\}_{i=1}^k$ (Задача классификации на k классов).

А также наивное предположение, которое делает байесовский классификатор наивным: все признаки независимые.

Согласно теореме Байеса,

$$P(y_i = C_j | x_i) = \frac{P(x_i | y_i = C_j) \cdot P(y_i = C_j)}{P(x_i)} \quad (1)$$

Так как все признаки считаем независимыми, то происходит следующая факторизация:

$$P(x_i | y_i = C_j) = \prod_{l=1}^p P(x_i^l | y_i = C_j)$$

Таким образом вероятность $P(x_i | y_i = C_j)$ мы поняли как считать. Вероятность $P(y_i = C_j)$ мы сможем посчитать по частоте встречаемости экземпляров класса в тренировочной выборке.

Так как у нас стоит задача классификации, то ответом к ней будет класс, к которому относится объект: $C = \arg \max_j P(y_i = C_j | x_i)$ - аргумент, при котором вероятность максимальна. Заметим, что максимизируем мы по различным классам, а знаменатель выражения (1) не зависит от j , поэтому для задачи максимизации он нам не нужен и мы его просто выбросим из рассмотрения. При желании его посчитать, можно воспользоваться формулой полной вероятности.

Градиентная оптимизация

Пусть модель зависит от параметров w : $f_w(X)$. Наша задача в повышении качества нейронной сети. Математически это качество выражается в виде функции качества, или обратной к ней - функции потерь. Т.о. наша задача состоит в минимизации функции потерь.

Для этого подберём параметры модели, при которых функция потерь будет наименьшей (**это и есть оптимизация**). Так как градиент - это направление наискорейшего роста, то будет подбирать параметры модели следующим образом:

$$w_{n+1} = w_n - \alpha \nabla_w L(y, f_w(X))$$

где (w_n) - последовательность параметров модели (при $n \rightarrow \infty$ последовательность должна сходиться к локальному минимуму), w_0 - начальные значения параметров, α - шаг оптимизации/learning rate - некоторый коэффициент, который может меняться на каждом шагу.

Условная оптимизация

Условная оптимизация - это оптимизация параметров, на которые наложены ограничения.

Например, в нашей задаче есть n классов и в каждом из них какое-то количество объектов. Вероятность посмотреть на элемент $i_{\text{того}}$ класса равна p_i . Эти вероятности и будем считать параметрами.

На них есть понятное ограничение: $\sum_{i=1}^n p_i = 1$